

Tilburg University

Resources for mental lexicon research

Keuleers, Emmanuel; Marelli, Marco

Published in:
Word Knowledge and Word Usage

DOI:
[10.1515/9783110440577-005](https://doi.org/10.1515/9783110440577-005)

Publication date:
2020

Document Version
Peer reviewed version

[Link to publication in Tilburg University Research Portal](#)

Citation for published version (APA):
Keuleers, E., & Marelli, M. (2020). Resources for mental lexicon research: A delicate ecosystem. In V. Pirrelli, I. Plag, & W. U. Dressler (Eds.), *Word Knowledge and Word Usage: a Cross-disciplinary Guide to the Mental Lexicon* (pp. 167-188). De Gruyter. <https://doi.org/10.1515/9783110440577-005>

General rights

Copyright and moral rights for the publications made accessible in the public portal are retained by the authors and/or other copyright owners and it is a condition of accessing publications that users recognise and abide by the legal requirements associated with these rights.

- Users may download and print one copy of any publication from the public portal for the purpose of private study or research.
- You may not further distribute the material or use it for any profit-making activity or commercial gain
- You may freely distribute the URL identifying the publication in the public portal

Take down policy

If you believe that this document breaches copyright please contact us providing details, and we will remove access to the work immediately and investigate your claim.

Emmanuel Keuleers and Marco Marelli

Resources for mental lexicon research: A delicate ecosystem

Accepted Author Manuscript of:

Keuleers, E., & Marelli, M. (in press). Resources for mental lexicon research: A delicate ecosystem. In Vito Pirrelli, Ingo Plag, & Wolfgang U. Dressler (Eds.), *Word Knowledge and Word Usage: A Cross-disciplinary Guide to the Mental Lexicon*. Berlin: De Gruyter.

Abstract

Resources are playing an ever-increasing role in current empirical investigations of the mental lexicon. Notwithstanding their diffusion and widespread application, lexical resources are often taken at face value, and there are limited efforts to better understand the dynamics and implications subtending resource developments, as well as the complex web of relations linking resources to each other. In the present chapter, we argue that describing these dynamics and relations is akin to investigating a complex and delicate ecosystem: resources are not independent and self-contained elements, but are rather the expression of a set of entangled components that span from our everyday language experience to the very scientific theories we develop to understand language.

Emmanuel Keuleers
Department of Cognitive Science and Artificial Intelligence
Tilburg University
Warandelaan 2, 5037 AB Tilburg, The Netherlands
E.A.Keuleers@tilburguniversity.edu

Marco Marelli
Department of Psychology
University of Milano-Bicocca
Piazza dell'Ateneo Nuovo, 1
20126 Milano, Italy
marco.marelli@unimib.it

The authors would like to thank Dr. Paweł Manderer for many insightful comments on the topic of this paper.

1 Introduction

Knowledge about the nature and organization of the mental lexicon is strongly dependent on a large amount of resources, which, at first sight, seem relatively independent from each other. A first group of resources provides researchers with objective information on the elements that make up the lexicon in its different linguistic and psycholinguistic interpretations. There are data reflecting properties such as word length, morphology, or pronunciation; data concerning distributional properties of words based on text corpora; lexicographic data with definitions and relations; and so forth. A second group of resources are derived from behavioral or neuropsychological investigation using the elements of the lexicon as stimuli: subjective expressions of single word properties or word relatedness; response latencies; eye movement trajectories; encephalographic activity; etc. Finally, there are resources which inform researchers about more abstract properties of the lexicon and its elements, such as linguistic grammars, cognitive theories and formalisms, algorithms for lexical analysis and word tagging, etc.

In this chapter, we will show that gaining a proper understanding of these resources requires seeing them as part of a complex dynamic system. Figure 1 illustrates this dynamic view on psycholinguistic resources. In Figure 1, circles represent resource “primitives”, raw data that are used to develop the resources themselves. These primitives are restricted to language-associated human behavior. We conceive of these primitives as latent variables. They are not a directly accessible resource but instead they are the origin of the data we use. Every other linguistic or psycholinguistic resource can be considered a direct or indirect elaboration of these primitives. These general resource categories are represented as rectangles in Figure 1. The operation needed to derive a given resource type from another is represented by arrows. This schematic representation will serve as a guide throughout this chapter. We roughly divide our exposition in three parts. First, we will focus on the rightmost section of the Figure, highlighted in red. This part will deal with resources that are mostly related to psycholinguistics and cognitive linguistics, as it focuses on resources obtained through elicited behavior. Second, we will discuss resources that are mostly related to the collection of unelicited language behavior (leftmost part of

the Figure, in blue), and that are often developed in the computational domain. Finally, we will briefly consider the center of our representation, and argue that the very cognitive models and linguistic theories driving our research activity can be also considered resources that we use for investigating language. Note, however, that this subdivision is extremely rough. We have already mentioned that we believe that resources are not independent self-contained elements, but are rather the expression of a complex dynamic system that span from our everyday language experience to the very scientific theories we develop to understand language. This should be already evident by the deep entanglement between the elements we report in our figure: there is no isolated component, with most elements involved in ingoing and outgoing operations crucially binding them with each other. Indeed, given these considerations we believe that providing a complete taxonomy of resources is an impossible feat. The organization of this chapter has thus to be considered more of a working hypothesis, adopted for purely explanatory purposes.

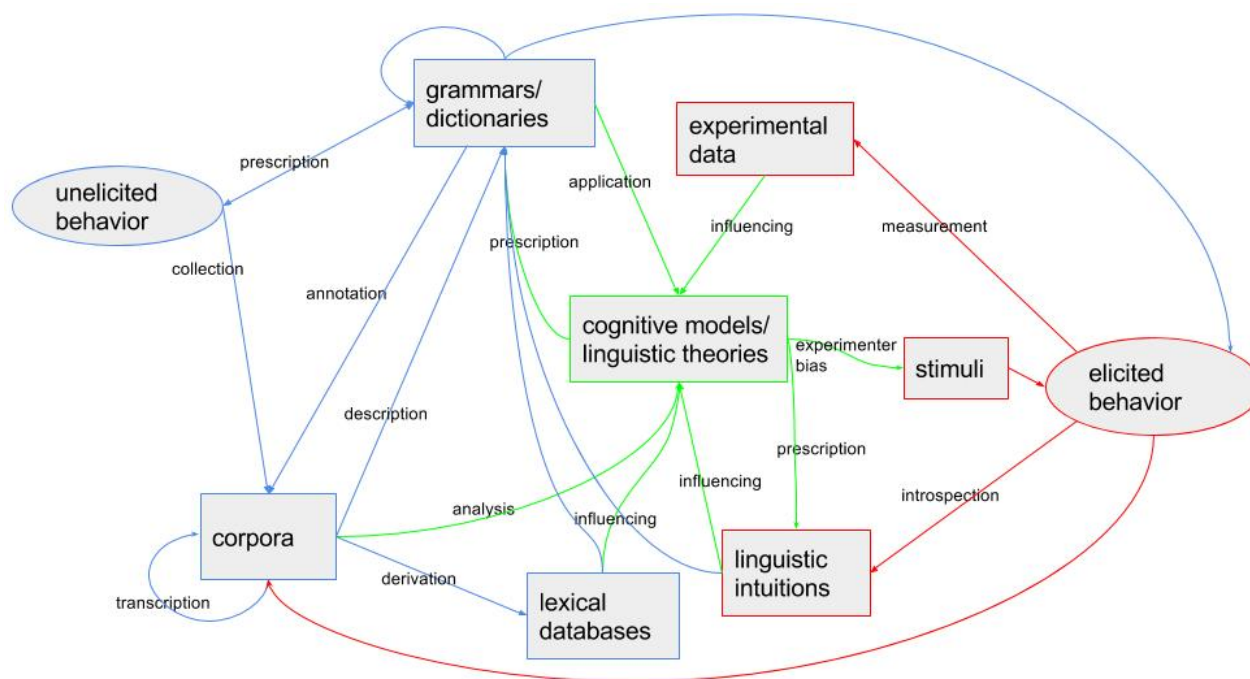


Figure 1: The resource ecosystem

2 From elicited behavior to experimental data and linguistic intuitions

In one way or another, human behavior underlies every form of linguistic data. In the present section we will focus on *elicited behavior*, or, in other words, behavior that is initiated at the researcher's input. In our schema (see Figure 1), the material that the researcher chooses to initiate a behavior is called *stimulus*. Depending on the researcher's intention, a stimulus will lead to *experimental data*, *linguistic intuitions*, or *corpora*. We will discuss the first two in this section, while corpora, which are most often not intentionally elicited, will be discussed in a later section on unelicited data.

It is important to note that *stimuli* are themselves a resource. Psycholinguistics has a long tradition in making data available to make it easier for researchers to select stimuli. An early example is found in Haagen (1949), who collected ratings for familiarity and vividness of imagery for more than 400 adjectives, in addition to ratings for synonymy and association of 440 adjective pairs. She specifically noted that the "study was undertaken to develop materials that might be used in studies of verbal learning" (Haagen 1949: 454). Similar efforts were conducted by Paivio, Yuille, and Madigan (1968) and by Nusbaum, Pisoni, and Davis (1984). The MRC database (Coltheart 1981) marks a radical change in the content of *stimulus* resources. While the purpose of the database remained unchanged – to serve as stimuli for psycholinguistic experiments – elicited ratings for words were merged with data from other resources containing information about syntactic class, formality of register, pronunciation, and word frequency. We find these latter resources scattered throughout our schema – a first indication of the dynamic nature of the resource system. Another class of *stimulus* resource are collections of pseudowords, which are essential to tap into the productive aspect of language. While the ARC database (Rastle, Harrington, and Coltheart 2002) still exhaustively lists pseudowords, subsequent efforts at building pseudoword resources, such as WordGen (Duyck et al. 2004), LiNGUA (Westbury, Hollis, and Shaoul 2007), and Wuggy (Keuleers and Brysbaert 2010) abandoned the listing approach in favor of algorithms that could dynamically generate pseudowords with certain properties. Following the theme that will pervade this chapter, these *stimulus* resources rely heavily on other resources,

primarily on lists of existing words taken from *lexical databases*. While these resources are practical tools for researchers, the most frequent *stimulus* resources are probably lists of stimuli used in previous research and often appearing in the appendices of published papers, especially when it comes to the investigation of rare phenomena. In the domain of morphology, for example, lists of opaque words are often reused for the investigation of semantic processing. As noted above, *stimulus* resources are rarely limited to orthographic or phonetic strings. Most often they also merge estimates of different properties of these strings. In that respect, they have a certain degree of overlap with resources such as *dictionaries*, experimental *data*, and *frequency lists*. However, they crucially differ from these other resources in having a constrained use, namely to generate lists of stimuli for an experiment. Throughout the chapter, we will meet other examples of such apparent links, in which resources quite similar in substance are developed for and applied to different purposes, time after time illustrating an intertwined, dynamic, and complex system.

2.1 Experimental data

For the present purpose, we define *experimental data* as the result of measuring the response to stimuli using objective instruments. Measurements in psycholinguistics are usually chronometric or physiological. Chronometric measures are the result of recording elapsed time, for instance the measurement of reaction time in a word identification task or the measurement of fixation durations during reading. Physiological measures are the result of recording electrical (EEG) and magnetic (MEG) signals generated by the brain, recording the change in blood oxygenation level in grey matter (fMRI), or recording more peripheral activity, such as skin conductance, electrical activation in muscles, or pupil dilation.

Experimental data are often published in the normal course of research and are undeniably an exploitable resource. Meta-analysis can be used to combine data from several similar experiments to increase the strength of a statistical analysis; computational models can be validated by experimental data, etc. Unlike *stimulus* resources, experimental data are usually collected with the aim of testing specific hypotheses. Their use as a resource is mostly *secondary* and coincidental.

2.2 Megastudies

Megastudies are a category of experimental data whose *primary* purpose is to function as a resource. These data are collected specifically with the aim of maximizing utility and re-usability in the context of psycholinguistic research. In this aspect, they are similar to databases of ratings, but they differ firstly in the sense that what is being collected is measured via objective instruments and secondly in the sense that the collected measurements are usually considered to be *dependent* variables in experimental research. While experimental psychologists have long been committed to building and using resources of independent variables, such as the stimulus resources discussed above, they have been reluctant to build large collections of responses to those stimuli. In fact, the earliest collection of chronometric data that was designed specifically with re-use in mind (Balota et al. 2007) was published more than 60 years after Haagen's (1949) collection of stimulus ratings. Keuleers and Balota (2015) have tried to explain this time gap by a dogmatic adherence to strict temporality in the cycle of experimental research.

"In hindsight, one can ask why the psychologists who understood the benefit of collecting elicited ratings for a large number of words did not gather chronometric measures for recognition or classification of those words. One possibility is that the reuse of independent variables was considered safe but that recycling a dependent variable did not conform to the idea that formulating a hypothesis must always precede the collection of data in the cycle of scientific investigation. The fundamental idea behind that principle, however, is to prevent a hypothesis being generated based on data that are already collected. It is clear to see that a careless generalization of precedence in the scientific cycle to precedence in time is absurd, as it would imply that temporally earlier findings cannot be used to contest the validity of later findings." (Keuleers and Balota, 2015:1459)

In line with this interpretation, it took some serendipity for psychologists to realize that objective dependent measures such as reaction times could be collected with reuse in mind. The events leading to this accidental insight started with Seidenberg and Waters (1989) who collected voice onset times for about 3,000

monosyllabic English words at McGill University. Their purpose was to compare the amount of variance in naming latency that could be explained by different theories of reading aloud. Seidenberg and Waters coined the term *megastudy* to refer to the – for that time – unusually large number of stimuli. With an entirely different purpose, Treiman et al. (1995) re-used the McGill dataset to test hypotheses on the role of spelling consistency in reading aloud. In doing so, they implicitly acknowledged that an existing dataset could be used to examine a novel research question. However, they still seemed to consider the McGill dataset as merely a source of supporting evidence for results they had already obtained in their own experiments.

A few other studies followed, using more or less the same sets of items. Spieler and Balota (1997, 2000) collected naming times for both younger and older adults; Balota et al. (2004) did the same using lexical decision instead of naming.

The revolution in megastudy data came with the publication of the English Lexicon project (Balota et al. 2007), which provided both lexical decision and naming data for more than 40,000 words, collected at six different universities. The authors of the English Lexicon project were clear in their motivations: the database was to be used as a normative standard for lexical decision and naming in English. This would free researchers from the need to do a plethora of small factorial experiments in their laboratories, instead enabling them to look at the functional relationship between their variables of interest and visual word processing data.

The English Lexicon project was soon followed by the French Lexicon project (Ferrand et al. 2010), which collected lexical decision data for nearly 40,000 words, and some smaller efforts such as the Malay lexicon project (Yap et al. 2010). Keuleers, Diependaele and Brysbaert (2010) made the process of collecting megastudy data much more efficient, by demonstrating that individual participants can yield reliable data for tens of thousands of trials. This approach resulted in the Dutch Lexicon project (Keuleers, Diependaele & Brysbaert, 2010; Brysbaert et al. 2016) and the British Lexicon project (Keuleers et al. 2012).

The megastudy approach was also quickly extended from simple visual word recognition to other, more complex paradigms at the word level, such as semantic priming (Hutchison et al., 2013), masked priming (Adelman et al. 2014), auditory lexical decision (Ernestus and Cutler 2015) and recognition memory (Cortese, Khanna, and Hacker 2010; Cortese, McCarty, and Shock 2015). More recently, large datasets of reading at the sentence level, such as the GECO eye-tracking corpus, have also become available (Cop et al. 2016).

Given the success of megastudy resources, researchers did not mind advancing knowledge from existing experimental data: megastudy data or the studies that were based on them data were not criticized because they violated the temporal precedence of hypothesis generation to data collection. Still, as Keuleers and Balota (2015) have pointed out, when data are available before the hypotheses are formulated, there is a real danger of data-driven hypothesis generation. Fortunately, researchers have started to address this problem by using methods such as bootstrapping (Kuperman 2015).

2.3 Clinical resources

Similar in concept to megastudies are resources that bundle experimental data from patients with language-related clinical symptoms. The Moss Aphasia Psycholinguistics Project Database (Mirman et al. 2010) contains picture naming and picture recognition data for 175 items from over 240 patients. For many of these patients, there are also demographic data, aphasia diagnostic tests, speech perception and recognition tests, and a variety of other language and memory tests. A more in-depth overview of large datasets in clinical research is provided by Farooqi-Shah (2016).

2.4 Crowdsourcing

While researchers are now mostly convinced that data collected on thousands of items can yield valid data for scientific investigation, another hurdle to overcome is the idea that reliable data can only be collected in controlled laboratory

experiments. In the context of psycholinguistic research, crowdsourcing is used when elicited data (experimental data or intuitions) are collected outside laboratory settings from a large set of participants whose demographic characteristics are not known a priori. Recently, however, researchers have used crowdsourcing to create resources collected on very large and diverse samples. In the context of visual word recognition, Dufau et al. (2011) have started an effort using a specialized mobile app to generate lexical decision data in different languages. More recently, attention has shifted to collecting data by offering participants a game-like format to test their vocabulary. This has resulted in large resources containing data about word knowledge and word recognition times for over 50,000 Dutch words collected on several hundred thousand participants (Keuleers et al. 2015), for over 60,000 English words collected on over a million participants (Brysbaert et al. in press), and for over 45,000 Spanish words collected on over 160,000 participants (Aguasvivas et al., 2018)

An essential aspect of crowdsourcing in science is that part of the work of the scientist is transferred to laypersons, who each contribute a small part of the data. It could be argued that crowdsourcing has been an integral method in psycholinguistics from very early on because, unlike in other sciences where a skilled scientist who is familiar with an instrument can make better observations than a layperson, psychological observations are dependent on the naivety of the respondent, because involvement with the goals of the research would taint the results.

2.5 Linguistic Intuitions

A critical aspect of experimental data is the involvement of objective measurement. However, behavior can also be elicited by a researcher with the aim of self examination. This is common in theoretical linguistics, for instance in the case of grammaticality judgements. The same approach is used in the compilation of psycholinguistic resources that rely on questionnaires asking participants about their intuition on certain aspects of linguistic experience, such as word age of acquisition

(e.g. Kuperman, Stadthagen-Gonzalez, and Brysbaert 2012), valence, dominance, and arousal (Warriner, Brysbaert, and Kuperman 2013), concreteness (Brysbaert, Warriner, and Kuperman 2014), modality-specificity (Lynott and Connell 2013), or semantic features (McRae et al. 2005). The two critical differences between the results of the questionnaires that psycholinguists administer and the intuitions that theoretical linguists supply is that the data from questionnaires are *aggregated* over multiple participants and that the participants are *naïve*. Thus, it is clear that when grammaticality ratings are collected on naïve participants and aggregated (e.g., Bard, Robertson and Sorace 2015), there is no difference between the two.

The notion that a linguistic intuition is an (self-)elicited response simply means that theoretical linguists administer themselves examples of language usage as *stimuli* in order to produce the intuitions (*or responses*) that are at the center of their research. The terminology of stimulus-response is closely connected with behaviorism and therefore seems irreconcilable with the views espoused in generative grammar which use linguistic intuitions as a primary resource. It should be clear, however, that using a stimulus-response based research paradigm to gather data does not imply that the faculty of language operates on behaviorist principles. In the context of this work, the terminology allows us to consider both ratings and intuitions as closely related psycholinguistic resources.

Another important point is that elicited responses can result simultaneously in both experimental data and linguistic intuitions. For instance, in the lexical decision paradigm, the decision whether a stimulus is a word or not is an intuition or judgement (and indeed it is usually released in the context of related resources; e.g., Keuleers et al. 2015), whereas the time taken to make that decision is experimental data. To make the difference between these concepts even more clear: If we were to ask participants to estimate the time it would take them to recognize a stimulus word, we would consider it an intuition. Likewise, asking participants to rate a particular aspect of a linguistic stimulus on a numerical scale is not an objective measurement but the recording of an intuition.

In the context of stimulus resources we have already discussed collections of ratings. These are obviously collections of linguistic judgements, but their primary use is as *a resource* for selecting stimuli and to function as an independent variable. Secondly, these ratings can also be treated, as described in the present section, as *dependent* variables providing inferential evidence for the development of cognitive models and linguistic theories.

3 From unelicited behavior to corpora and lexical statistics

Only an infinitesimal fraction of language production is elicited by scientists. Because language production is ephemeral, capturing it is notably difficult. Traditionally, language production was captured in field studies, providing direct access to language production. Still, even when there were direct means of recording the data, for instance through transcriptions, this was mostly limited to an extremely small fraction of the full range of language experiences. Cultural and historical changes have made this endeavor progressively more feasible. Increasing literacy in the general population, and the evolution of printing techniques first caused a massive growth in the production and availability of written language. Then, the development of audio and video recording made it possible to extend data collection to spoken data and gestures. Finally, the digital revolution had such an influence on the development of linguistic resources that nowadays we cannot imagine a non-digital corpus. Digital technologies are helping to collect and store progressively larger amount of language production. Communication networks have also made the dissemination of the resources much faster. In addition, the digital world has become a source of peculiar language data and investigating the language used in social media and the web is now a central topic of study (e.g., Schwartz et al., 2013; Herdağdelen and Marelli, 2017).

While this progress has had an obvious impact on the development of various, large corpora, we should not forget that the evolution of grammar and lexicography is also a result of the technological innovations that make the examination of language possible independent of space and time. It is no exaggeration to say that these changes have revolutionized the way language science is done.

The present section will focus on the linguistic resources that are prominently based on unelicited behavior. Most notably, we will focus on corpora, lexical databases, and dictionaries and grammars.

3.1 Corpora

A corpus can be defined as a collection or sample of language events that are related to each other in one or more aspects. For example, the events can have the same source (e.g. newspaper writings, books, dialogue, etc.) or modality (e.g. written text, speech, gestures, video-recordings).

As mentioned earlier, corpus is now mostly used as synonym for digital corpus. However, this relatively recent trend can be traced back to the 1950s, with Padre Busa's "Index Thomisticus", an annotated and lemmatized corpus of the works by Thomas Aquinas. Another milestone in modern corpus linguistics came with the publication of "A Computational Analysis of Present-Day American English" (Kučera and Francis 1967), also known as the Brown corpus. This resource is still quite popular in many domains, notwithstanding its now well-known shortcomings (see below). Today, the size of these pioneering collections looks extremely limited. During the last two decades, we have seen a massive increase in the average corpus size, with modern corpora often containing billions of tokens. This rapid growth in the size of corpora is strictly related to the increasingly stronger association between computational linguistics and the web, that represents a massive, always-growing, and easy-to-harvest source of language data.

One result of this is that modern corpora will no longer be considered as static but as dynamic sources, where content is added on a real-time basis. One example of this is the NOW news corpus (Davies 2013), which currently contains 4.5 billion words and grows at a rate of nearly 6 million words each day. In other words, the modern conception of corpus as a very large digital collection of language material is bound to be superseded by the conception of a corpus as a stream.

The digital revolution also had profound repercussions on the treatment and processing of corpora. Not only has digitization made text processing much faster, it also has increased the synergy between corpus linguistics and resources from other domains. For instance, it has become trivial to annotate a text corpus with any information about a word found in dictionaries or other lexical databases, thanks to tools from natural language processing (e.g. Part-of-Speech taggers, lemmatizer, and parser). However, while these resources have made corpus annotation easier, they have also brought with them an unavoidable imprecision in the annotation itself. No automatic annotation is perfect, and formal evaluation in this respect is only reliable to a certain extent: the state-of-the-art of a given method depends on a comparison with a gold standard which may have an obscure origin, or may not fit well the specific purpose of a researcher. The application of NLP tools in the development of the corpus can have a massive influence on the corpus itself and on the research that is being done using the corpus. This warning should not be forgotten or underestimated: the influence of computational methods on linguistic resources is so profound that it quickly becomes impossible to disentangle effects of resources from effects of computational methods. From the moment that the behavior in a corpus is annotated using an automated method, the corpus as a linguistic resource becomes tainted by previous linguistic resources and taints subsequent resources. And from the moment a computational method is trained using corpus data, the subsequent application of the method to other data becomes tainted by the initial corpus data. These loops of cross-fertilization characterize the picture of language resources that we are drawing in the present chapter.

Corpora represent an ideal case study for the complex dynamics in resource ecosystem. There are mutual influences between different corpora and between corpora and other resource types. Modern corpora also lead to difficulties when trying to pigeonhole resources into strict taxonomies. Strictly speaking, a corpus is not necessarily a collection of samples of *unelicited behavior*. For instance, when behavior is recorded at the scientist's request, this request itself can influence the behavior, which should therefore be called partially elicited. While in its formal aspects the

behavior looks like unelicited behavior (e.g. unrestricted speech), researchers need to be aware of the ways in which the behavior may conform to the participants' expectations of the requirements of their behavior. A typical example in this respect is the CHILDES project (MacWhinney 2000), which contains many records of spontaneous mother-child interactions in a controlled environment, at the researcher's request. Related to CHILDES is the TalkBank project (MacWhinney 2007), a varied collection of resources, ranging from structured elicitation to free discourse data from typical and disordered populations. In language research, many corpora walk the fine line between elicited and unelicited of behavior.

3.2 Lexical databases

Entangled with corpora and computational methods in the resource ecosystem, we find lexical databases: collections of words that have been associated with one or more word properties. The properties are often derived from corpora, but can also be derived from other databases, experimental data, or other resources. Lexical databases can span from relatively simple resources, such as frequency norms, to data obtained through complex computational systems, such as automatically-obtained word meaning relations. In one of the typical loops of the resource ecosystem, lexical databases can also influence corpora, when they are used as a means for corpus annotation.

Frequency norms, or word occurrence counts, are the most typical example of a lexical database. Good frequency norms require high quality corpora, both in terms of size and source, and technologies to quickly process textual data. One of the first lexical databases containing word counts was developed as an educational resource. In the 1920s, Thorndike and Lorge started counting words occurring in magazines, newspapers and other contemporary sources with the express aim of providing a resource by which educators could select words for teaching materials. Their first publication was "A Teacher's word book of twenty thousand words" in 1931, followed by the thirty-thousand word version in 1944 (Thorndike and Lorge, 1944). This resource was almost immediately appropriated for scientific use most notably when

Howes and Solomon (1951) published their seminal study on the effect of word frequency on word identification speed. Like the Thorndike-Lorge norms, many word frequency lists developed in the 20th century were distributed in book form. Although some older frequency resources are still available in book form, one of the consequences of larger corpora is that word frequency lists also grow. While it does not take more space to increase the counter for a word that has already been encountered, each new word that is discovered requires extra space. As a result, the adoption of better frequency norms based on larger corpora was crucially dependent on the adoption of a digital approach and today's massive corpora have led to word frequency resources that are only digitally available. Digital storage has also made it possible to distribute frequencies for *n*-grams (sequences of *n* successive words). Although a text of 1000 words has 1000 single words and 999 bigram tokens, the bigrams are far less likely to occur multiple times than single words and therefore lists of *n*-grams are much larger. For instance, the SUBTLEX-UK word frequency list (van Heuven et al. 2014) contains counts for nearly 350,000 word and nearly six times as many bigrams. Besides word frequencies, other simple count measures include document and page counts, that form the basis for measures of diversity or dispersion. More specialized or rarely-used count measures are often computed when needed, rather than disseminated with the lexical database.

The influence of frequency norms on other resources is remarkable, especially in the resources based on unelicited behavior. Count data is often used as information in lexicographic work. For instance, the decision whether to include a word in a dictionary can be based on its frequency in recent texts. Likewise, the decision to remove a word from a practical rather than historical dictionary can be made on the basis a very low occurrence in contemporary word counts. Resources based on elicited behavior can also be affected by considerations related to lexical frequencies. In psycholinguistics, for instance, it is well known that word frequency has a large influence on language processing tasks. When resources such as stimulus lists are constructed, they often use frequency data as a guide to decide whether to include or exclude stimuli.

Counts can also act as the building block for more complex resources that aim at capturing higher-level linguistic information. For example, matrices that encode how often words are found together in a sentence or how often words are found in each document in a corpus form the basis of *vector space modelling*. These matrices, in which each word is represented by a series of numbers (vectors), can be processed through mathematical techniques in order to derive convenient data-driven representations of word meanings. This approach to semantics rests on the distributional hypothesis, stating that the meaning of a word can be approximated by the contexts in which that very word appears (Harris 1957), a general idea which traces back to philosophical proposals that are exemplified in Wittgenstein's works. The development of computational vector space modelling is relatively recent and makes use of such techniques as Latent Semantic Analysis (Landauer and Dumais 1997), Hyperspace Analogue to Language (Lund and Burgess 1996), and Latent Dirichlet Allocation (Blei et al. 2003). In such systems, semantic similarity is modelled in geometrical terms: since co-occurrence counts can be taken as coordinates in a high-dimensional space, the closer two vectors are, the more similar the corresponding word meanings will be. This is a direct consequence of the distributional hypothesis: words with similar meanings will often be found with the same surrounding words, leading to similar co-occurrence vectors. The approach was proven successful in capturing human intuitions concerning word meanings, and was then used as a way to automatically obtain semantic information in a number of domains, such as estimation of semantic relatedness and feature extraction. The approach is also used extensively in more applied natural language processing applications.

Psycholinguists often use the output of a distributional model as a resource for abstract semantic word representations. Two notably easy-to-access resources which, among other things, allow researchers to get distances between different semantic vectors are the LSA website (<http://lsa.colorado.edu>) and Snaut (Mandera, Keuleers, and Brysbaert 2017; available at <http://meshugga.ugent.be/snaut/>). More generally, there are a number of programming libraries available to build semantic vectors from a given corpus and to manipulate these semantic vector spaces, such

as word2vec (Mikolov et al. 2013), Dissect (Dinu, The Pham and Baroni 2013), Gensim (Řehůřek and Sojka 2011), and Tensorflow (Abadi et al. 2016).

While these techniques usually take unelicited behavior as their input, they are in fact completely agnostic to the origin of the co-occurrence data. For instance, Andrews, Vinson, and Vigliocco (2009) developed a multimodal distributional model that combines text-based data and human-generated experiential information, and De Deyne, Verheyen, and Storms (2015) have developed systems based on relatedness judgements. Moreover, in principle the techniques can work on any input modality, so that gesture-based models, sounds, and images can also be processed in similar ways.

It is however evident that quantitative representations for words, whether they are simple word frequencies or more complex estimates, are greatly influenced by the corpora that they are based on. In a very broad sense, the “world” that is captured by the corpus will also transpire in the derived measures: you can take the word out of the corpus but you can’t take the corpus out of the word representation. Indeed, Louwerse and Zwaan (2009) have shown that the precision of text-based geographical estimates is associated with the physical distance between the text source and the considered place: the NY Times is better suited at estimating the location of East-Coast cities, and the LA Times is better suited at estimating the location of West-Coast cities. As a consequence, quantitative representations can not be regarded as unbiased samples of behavior, but should rather be always interpreted with their provenance in mind.

Lexical databases can also encode non-quantitative properties of words. A prominent example is WordNet (Miller et al. 1990), a large database that has been extremely influential in both cognitive science and computational linguistics. WordNet can serve as a collection of word senses, listing all possible meanings that a word can denote. In addition, WordNet provides information about synsets (groups of word that are pure synonyms), as well as different types of semantic relations between words (e.g. oak is a hyponym of tree). In the context of the resource ecosystem,

Wordnet is a very strange beast. While it has all the characteristics of a lexical database, it is also extremely close to being a dictionary and a thesaurus (see below). Moreover, it is developed with an explicit reference to cognitive models of human semantic memory, making it a good candidate for what we called linguistic-intuition resources: in WordNet, words can be seen as self-administered stimuli for which experts provide their educated opinion. Additionally, even if such a claim was not advanced by its proponents, in computational linguistics WordNet is often considered an ontology, that is, a resource encoding the types, properties, and interrelationships of entities in the world.

WordNet can be taken as a further example of the entanglement between the components of the resource ecosystem. It results from the combination of several techniques used for resource development and illustrates the weak boundaries of different resource types when a rigid resource taxonomy is used. Practically, WordNet is often used as data source for techniques that are in turn at the basis of the development of other resources. Most notably, WordNet is a popular resource for the estimation of word meaning similarity, making it a primary influence on other lexical databases.

3.3 Dictionaries and grammars

The goal of recording vocabulary is very old. Dictionaries certainly predate the dramatic amounts of recorded language behavior that is available today. Lexicographic work is traditionally the result of observing *unelicited behavior*. However, the resource ecosystem reveals a more complex picture. An essential step in any lexicographic enterprise consists of listing words in a language, which may also be based on the examination of recorded behavior with the aim of discovering new word types. This means that, in the massively inter-influencing resource system, dictionaries and grammars are largely influenced by corpora. In the late 20th century, dictionary makers such as Collins started building corpora with the explicit aim of identifying new words and informing dictionary development (Sinclair 1987).

Development of dictionaries is almost always driven by other dictionaries and grammars, as they are almost never written without support from earlier resources of the same type. It can be argued that while *listing* the words is based mostly on unelicited behavior, every other aspect of lexicography mostly consists of self-elicited behavior equivalent to the linguistic intuitions we discussed earlier (for instance: definitions, lexical and ontological relationships), making the boundaries between dictionaries and other resources even fuzzier.

Dictionaries and other word lists are extremely influential as a linguistic resource. Because they are a reflection and a source of authority on the use and meaning of words, they modulate any type of human linguistic behavior, either elicited or unelicited. It could be said that of all linguistic resources, dictionaries influence language behavior the most. We could even ask the question whether language behavior influences dictionaries more than dictionaries influence the behavior itself.

Next to the recording of words, the recording of how words are used in different contexts and how they combine together in sentences (i.e. grammar) is one of the earliest linguistic endeavors. Rather than exhaustively listing, which is the goal of a dictionary, the goal of a descriptive grammar is to compress knowledge. Concepts like conjugation, inflection, syntactic classes, sentences and clauses allow for substituting lists of individual instances for a description of rules and exceptions. Like dictionaries, grammars influence behavior from the moment they exist and the more authority they receive, the more they influence the behavior.

For the reasons described above, relying on other resources for the construction of dictionaries and grammars leads to a cycle of self-reinforcement of prescriptive language behavior. This is not only the result of the obvious influence of pre-existing dictionaries, but also of the corpora themselves. It is in fact important to note that unelicited behavior is not equivalent to non-scripted behavior. Recorded behavior is very much biased to highly edited and scripted production. Written behavior is almost never captured before it has gone through several stages of editing

(indeed, there are some specific corpora dealing with capturing the editing process itself, for instance in research on journalism). Spoken language production, especially the examples that can be found in corpora, is neither necessarily unscripted (e.g. films and tv programs in subtitle corpora). As a result, a large part of the language behavior considered in lexicography is already implicitly adherent to the prescriptive rules imposed by dictionaries and grammars.

This does not mean that all linguistic behavior is influenced by prescriptive resources. However, we should be aware that where editing and scripting are involved, the prescriptive influence is probably strong. This tendency becomes even more pronounced as the editing phase in language production becomes more and more driven by artificial software that directly interfaces with the resources. Consider how spellcheck and grammar check determine our online behavior in written production or how personalized dialogue systems (such as Apple Siri, Amazon Echo, or Google Home) recognize some commands while they do not recognize others. As a consequence, the connection between prescriptive sources and production get stronger with time, with technological innovation as its catalyst. On the other hand, it is also true that the massive availability of unelicited behavior makes the inclusion of new words or constructions more probable.

4 Cognitive models and linguistic theories: Feedback at the core

Up to this point, we have tried to frame resources in an atheoretical way. However, as Figure 1 reveals, theories and models are at the center of our formalization. They occupy the box with the largest number of connections, with outgoing arrows showing that theories heavily influence resource development and incoming arrows representing how theories are developed on the basis of the available resources.

This entanglement has far reaching implications for the epistemological status of resources and cognitive models or linguistic theories. It cannot be said that any of them are independent and contain an objective truth. The only exception to this

would be the observation of unelicited behavior in a group of language users who have no concept of linguistic resources. In other words, when language is used in a context without any resource, its behavior can be regarded as unbiased. On the other extreme, there are such languages as modern English, where it has become impossible to disentangle the language behavior from the influence of the resources. Child language is no exception to this as it is completely contingent on the language of adults, which itself a product of interaction between resources and behavior.

In this light, it is important to understand that any cognitive model or linguistic theory that is informed by such a *cultivated and resource-driven* language must acknowledge this fact and its consequences. One of the more important consequences is that certain aspects of language behavior may only arise in resource driven languages and not in language in its “ideal” pre-resource state. In other words, neither the language behavior nor the language faculty that we can observe today should be regarded as emerging from the simple interaction between humans endowed with the capacity for speech. Instead, we should always keep in mind that resources shape language, and that there is a constant feedback between language behavior and its resources. This entanglement will only become more pronounced as technological innovations become more related to the production of language. As a simple example, predictive text input, which is of course based on algorithms that interface with linguistic resources, influences language behavior at the exact moment it takes place. Technologies like grammar and spell-checking are also instances of this extreme entanglement between resources and language production.

The influence that linguistic theories have on resources derived from elicited behavior must also be acknowledged. However, this is perhaps less severe because it is epistemologically charted territory. It can suffice to classify this under experimenter bias and remedies to this bias are well known: responses should be elicited in double blind conditions so as to eliminate both experimenter bias and expectation bias. Scientists who wish to base their theories on their own intuitions should be aware of biases and strive to eliminate them or acknowledge that the subject under study is a language that is not only cultivated by prescriptive resources, but also by constraints on what the linguistic theory allows for.

5. References

- Abadi, Martín, Paul Barham, Jianmin Chen, Zhifeng Chen, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Geoffrey Irving & Michael Isard. 2016. Tensorflow: A system for large-scale machine learning. *12th USENIX Symposium on Operating Systems Design and Implementation (OSDI 16)*, 265–283.
- Adelman, James S., Rebecca L. Johnson, Samantha F. McCormick, Meredith McKague, Sachiko Kinoshita, Jeffrey S. Bowers, Jason R. Perry, et al. 2014. A behavioral database for masked form priming. *Behavior Research Methods* 46(4). 1052–1067. doi:[10.3758/s13428-013-0442-y](https://doi.org/10.3758/s13428-013-0442-y).
- Aguasvivas, Jose Armando, Manuel Carreiras, Marc Brysbaert, Paweł Mander, Emmanuel Keuleers & Jon Andoni Duñabeitia. 2018. SPALEX: A Spanish Lexical Decision Database From a Massive Online Data Collection. *Frontiers in Psychology* 9. doi:[10.3389/fpsyg.2018.02156](https://doi.org/10.3389/fpsyg.2018.02156).
- Andrews, Mark, Gabriella Vigliocco & David Vinson. 2009. Integrating experiential and distributional data to learn semantic representations. *Psychological Review* 116(3). 463.
- Balota, David A., Andrew J. Aschenbrenner & Melvin J. Yap. 2013. Additive effects of word frequency and stimulus quality: The influence of trial history and data transformations. *Journal of Experimental Psychology: Learning, Memory, and Cognition* 39(5). 1563–1571. doi:[10.1037/a0032186](https://doi.org/10.1037/a0032186).
- Balota, David A., Michael J. Cortese, Susan D. Sergent-Marshall, Daniel H. Spieler & Melvin J. Yap. 2004. Visual Word Recognition of Single-Syllable Words. *Journal of Experimental Psychology: General* 133(2). 283–316. doi:[10.1037/0096-3445.133.2.283](https://doi.org/10.1037/0096-3445.133.2.283).
- Balota, David A., Melvin J. Yap, Keith A. Hutchison, Michael J. Cortese, Brett Kessler, Bjorn Loftis, James H. Neely, Douglas L. Nelson, Greg B. Simpson & Rebecca Treiman. 2007. The English lexicon project. *Behavior Research Methods* 39(3). 445–459.
- Bard, Ellen Gurman, Dan Robertson & Antonella Sorace. 1996. Magnitude estimation of linguistic acceptability. *Language* 32–68.
- Blei, David M., Andrew Y. Ng & Michael I. Jordan. 2003. Latent Dirichlet allocation. *Journal of Machine Learning Research* 3(Jan). 993–1022.

- Brysbaert, Marc, Paweł Mander, Samantha F. McCormick & Emmanuel Keuleers. In press. Word prevalence norms for 62,000 English lemmas. *Behavior Research Methods*.
- Brysbaert, Marc, Michaël Stevens, Paweł Mander & Emmanuel Keuleers. 2016. The impact of word prevalence on lexical decision times: Evidence from the Dutch Lexicon Project 2. *Journal of Experimental Psychology: Human Perception and Performance* 42(3). 441–458. doi:[10.1037/xhp0000159](https://doi.org/10.1037/xhp0000159).
- Brysbaert, Marc, Amy Beth Warriner & Victor Kuperman. 2014. Concreteness ratings for 40 thousand generally known English word lemmas. *Behavior Research Methods* 46(3). 904–911.
- Coltheart, Max. 1981. The MRC psycholinguistic database. *The Quarterly Journal of Experimental Psychology Section A* 33(4). 497–505. doi:[10.1080/14640748108400805](https://doi.org/10.1080/14640748108400805).
- Cop, Uschi, Nicolas Dirix, Denis Drieghe & Wouter Duyck. 2016. Presenting GECO: An eyetracking corpus of monolingual and bilingual sentence reading. *Behavior Research Methods* 1–14.
- Cortese, Michael J., Maya M. Khanna & Sarah Hacker. 2010. Recognition memory for 2,578 monosyllabic words. *Memory* 18(6). 595–609.
- Cortese, Michael J., Daniel P. McCarty & Jocelyn Schock. 2015. A mega recognition memory study of 2897 disyllabic words. *The Quarterly Journal of Experimental Psychology* 68(8). 1489–1501. doi:[10.1080/17470218.2014.945096](https://doi.org/10.1080/17470218.2014.945096).
- Davies, Mark. 2013. Corpus of News on the Web (NOW): 3+ billion words from 20 countries, updated every day. URL <http://corpus.byu.edu/now>.
- De Deyne, Simon, Steven Verheyen & Gert Storms. 2015. The role of corpus size and syntax in deriving lexico-semantic representations for a wide range of concepts. *The Quarterly Journal of Experimental Psychology* 68(8). 1643–1664. doi:[10.1080/17470218.2014.994098](https://doi.org/10.1080/17470218.2014.994098).
- Dinu, Georgiana, Nghia The Pham & Marco Baroni. 2013. DISSECT - DIStributional SEmantics Composition Toolkit. *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, 31–36.
- Dufau, Stephane, Jon Andoni Duñabeitia, Carmen Moret-Tatay, Aileen McGonigal, David Peeters, F.-Xavier Alario, David A. Balota, et al. 2011. Smart Phone, Smart

Science: How the Use of Smartphones Can Revolutionize Research in Cognitive Science. (Ed.) Kenji Hashimoto. *PLoS ONE* 6(9). e24974.

doi:[10.1371/journal.pone.0024974](https://doi.org/10.1371/journal.pone.0024974).

Duyck, Wouter, Timothy Desmet, Lieven PC Verbeke & Marc Brysbaert. 2004. WordGen: A tool for word selection and nonword generation in Dutch, English, German, and French. *Behavior Research Methods, Instruments, & Computers* 36(3). 488–499.

Ernestus, Mirjam & Anne Cutler. 2015. BALDEY: A database of auditory lexical decisions. *The Quarterly Journal of Experimental Psychology* 68(8). 1469–1488.

doi:[10.1080/17470218.2014.984730](https://doi.org/10.1080/17470218.2014.984730).

Farooqi-Shah, Yasmeen. 2016. The Rise of Big Data in Neurorehabilitation. *Seminars in Speech and Language* 37(01). 003–009. doi:[10.1055/s-0036-1572385](https://doi.org/10.1055/s-0036-1572385).

Ferrand, Ludovic, Boris New, Marc Brysbaert, Emmanuel Keuleers, Patrick Bonin, Alain Méot, Maria Augustinova & Christophe Pallier. 2010. The French Lexicon Project: Lexical decision data for 38,840 French words and 38,840 pseudowords. *Behavior Research Methods* 42(2). 488–496.

Haagen, C. Hess. 1949. Synonymity, vividness, familiarity, and association value ratings of 400 pairs of common adjectives. *Journal of Psychology* 27.

Harris, Zellig S. 1957. Co-occurrence and transformation in linguistic structure. *Language* 33(3). 283–340.

Herdağdelen, Amaç & Marco Marelli. 2017. Social media and language processing: How Facebook and Twitter provide the best frequency estimates for studying word recognition. *Cognitive Science* 41(4). 976–995.

Heuven, Walter J. B. van, Pawel Mandera, Emmanuel Keuleers & Marc Brysbaert. 2014. SUBTLEX-UK: A new and improved word frequency database for British English. *The Quarterly Journal of Experimental Psychology* 67(6). 1176–1190.

doi:[10.1080/17470218.2013.850521](https://doi.org/10.1080/17470218.2013.850521).

Howes, Davis H. & Richard L. Solomon. 1951. Visual duration threshold as a function of word-probability. *Journal of Experimental Psychology* 41(6). 401.

Hutchison, Keith A., David A. Balota, James H. Neely, Michael J. Cortese, Emily R. Cohen-Shikora, Chi-Shing Tse, Melvin J. Yap, Jesse J. Bengson, Dale Niemeyer & Erin Buchanan. 2013. The semantic priming project. *Behavior Research Methods* 45(4). 1099–1114. doi:[10.3758/s13428-012-0304-z](https://doi.org/10.3758/s13428-012-0304-z).

- Keuleers, Emmanuel & David A. Balota. 2015. Megastudies, crowdsourcing, and large datasets in psycholinguistics: An overview of recent developments. *The Quarterly Journal of Experimental Psychology* 68(8). 1457–1468.
doi:[10.1080/17470218.2015.1051065](https://doi.org/10.1080/17470218.2015.1051065).
- Keuleers, Emmanuel & Marc Brysbaert. 2010. Wuggy: A multilingual pseudoword generator. *Behavior Research Methods* 42(3). 627–633.
doi:[10.3758/BRM.42.3.627](https://doi.org/10.3758/BRM.42.3.627).
- Keuleers, Emmanuel, Kevin Diependaele & Marc Brysbaert. 2010. Practice Effects in Large-Scale Visual Word Recognition Studies: A Lexical Decision Study on 14,000 Dutch Mono- and Disyllabic Words and Nonwords. *Frontiers in Psychology* 1.
doi:[10.3389/fpsyg.2010.00174](https://doi.org/10.3389/fpsyg.2010.00174).
- Keuleers, Emmanuel, Paula Lacey, Kathleen Rastle & Marc Brysbaert. 2012. The British Lexicon Project: Lexical decision data for 28,730 monosyllabic and disyllabic English words. *Behavior Research Methods* 44(1). 287–304. doi:[10.3758/s13428-011-0118-4](https://doi.org/10.3758/s13428-011-0118-4).
- Keuleers, Emmanuel, Michaël Stevens, Paweł Mander & Marc Brysbaert. 2015. Word knowledge in the crowd: Measuring vocabulary size and word prevalence in a massive online experiment. *The Quarterly Journal of Experimental Psychology* 68(8). 1665–1692. doi:[10.1080/17470218.2015.1022560](https://doi.org/10.1080/17470218.2015.1022560).
- Kučera, Henry & Nelson Francis. 1967. *Computational analysis of present-day American English*. Providence: R.I.: Brown University Press.
- Kuperman, Victor. 2015. Virtual experiments in megastudies: A case study of language and emotion. *The Quarterly Journal of Experimental Psychology* 68(8). 1693–1710.
doi:[10.1080/17470218.2014.989865](https://doi.org/10.1080/17470218.2014.989865).
- Kuperman, Victor, Hans Stadthagen-Gonzalez & Marc Brysbaert. 2012. Age-of-acquisition ratings for 30,000 English words. *Behavior Research Methods* 44(4). 978–990.
- Landauer, Thomas K. & Susan T. Dumais. 1997. A solution to Plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological Review* 104(2). 211–240.
- Louwerse, Max M. & Rolf A. Zwaan. 2009. Language encodes geographical information. *Cognitive Science* 33(1). 51–73.

- Lund, Kevin & Curt Burgess. 1996. Producing high-dimensional semantic spaces from lexical co-occurrence. *Behavior Research Methods, Instruments, & Computers* 28(2). 203–208.
- Lynott, Dermot & Louise Connell. 2009. Modality exclusivity norms for 423 object properties. *Behavior Research Methods* 41(2). 558–564.
- MacWhinney, Brian. 2000. *The CHILDES Project: The Database*. Vol. 2. Mahwah, NJ: Lawrence Erlbaum Associates.
- MacWhinney, Brian. 2007. The Talkbank Project. In Joan C. Beal, Karen P. Corrigan & Hermann L. Moisl (eds.), *Creating and Digitizing Language Corpora*, 163–180. London: Palgrave Macmillan UK. doi:[10.1057/9780230223936_7](https://doi.org/10.1057/9780230223936_7).
- Mandera, Paweł, Emmanuel Keuleers & Marc Brysbaert. 2017. Explaining human performance in psycholinguistic tasks with models of semantic similarity based on prediction and counting: A review and empirical validation. *Journal of Memory and Language* 92. 57–78.
- McRae, Ken, George S. Cree, Mark S. Seidenberg & Chris McNorgan. 2005. Semantic feature production norms for a large set of living and nonliving things. *Behavior Research Methods* 37(4). 547–559.
- Mikolov, Tomas, Kai Chen, Greg Corrado & Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.
- Miller, George A., Richard Beckwith, Christiane Fellbaum, Derek Gross & Katherine J. Miller. 1990. Introduction to WordNet: An on-line lexical database. *International Journal of Lexicography* 3(4). 235–244.
- Mirman, Daniel, Ted J. Strauss, Adelyn Brecher, Grant M. Walker, Paula Sobel, Gary S. Dell & Myrna F. Schwartz. 2010. A large, searchable, web-based database of aphasic performance on picture naming and other tests of cognitive function. *Cognitive Neuropsychology* 27(6). 495–504. doi:[10.1080/02643294.2011.574112](https://doi.org/10.1080/02643294.2011.574112).
- Nusbaum, Howard C., David B. Pisoni & Christopher K. Davis. 1984. Sizing up the Hoosier mental lexicon: Measuring the familiarity of 20,000 words. *Research on Speech Perception Progress Report* 10(10). 357–376.
- Paivio, Allan, John C. Yuille & Stephen A. Madigan. 1968. Concreteness, imagery, and meaningfulness values for 925 nouns. *Journal of Experimental Psychology* 76(1, part 2). 1–25.

- Rastle, Kathleen, Jonathan Harrington & Max Coltheart. 2002. 358,534 nonwords: The ARC nonword database. *The Quarterly Journal of Experimental Psychology: Section A* 55(4). 1339–1362.
- Řehůřek, Radim & Petr Sojka. 2011. Gensim—Statistical Semantics in Python. <http://www.fi.muni.cz/usr/sojka/posters/rehurek-sojka-scipy2011.pdf> (30 May, 2017).
- Schwartz, H. Andrew, Johannes C. Eichstaedt, Margaret L. Kern, Lukasz Dziurzynski, Stephanie M. Ramones, Megha Agrawal, Achal Shah, Michal Kosinski, David Stillwell & Martin EP Seligman. 2013. Personality, gender, and age in the language of social media: The open-vocabulary approach. *PloS One* 8(9). e73791.
- Seidenberg, Mark S. & G. S. Waters. 1989. Word recognition and naming: A mega study. *Bulletin of the Psychonomic Society* 27. 489.
- Sinclair, John M. 1987. *Looking Up: An Account of the COBUILD Project in Lexical Computing and the Development of the Collins COBUILD English Language Dictionary*. London: Collins ELT.
- Spieler, Daniel H. & David A. Balota. 1997. Bringing computational models of word naming down to the item level. *Psychological Science* 411–416.
- Spieler, Daniel H. & David A. Balota. 2000. Factors influencing word naming in younger and older adults. *Psychology and Aging* 15(2). 225–231. doi:[10.1037//0882-7974.15.2.225](https://doi.org/10.1037//0882-7974.15.2.225).
- Thorndike, Edward L. & Irving Lorge. 1931. *A teacher's word book of twenty thousand words*. New York: Teachers College, Columbia University.
- Thorndike, Edward L. & Irving Lorge. 1944. *The teacher's word book of 30,000 words*. New York: Teachers College Press.
- Treiman, Rebecca, John Mullennix, Ranka Bijeljic-Babic & E. Daylene Richmond-Welty. 1995. The special role of rimes in the description, use, and acquisition of English orthography. *Journal of Experimental Psychology: General* 124(2). 107.
- Warriner, Amy Beth, Victor Kuperman & Marc Brysbaert. 2013. Norms of valence, arousal, and dominance for 13,915 English lemmas. *Behavior Research Methods* 45(4). 1191–1207. doi:[10.3758/s13428-012-0314-x](https://doi.org/10.3758/s13428-012-0314-x).

Westbury, Chris, Geoff Hollis & Cyrus Shaoul. 2007. LINGUA: the language-independent neighbourhood generator of the University of Alberta. *The Mental Lexicon* 2(2). 271–284.

Yap, Melvin J., Susan J. Rickard Liow, Sajlia Binte Jalil & Siti Syuhada Binte Faizal. 2010. The Malay Lexicon Project: A database of lexical statistics for 9,592 words. *Behavior Research Methods* 42(4). 992–1003. doi:[10.3758/BRM.42.4.992](https://doi.org/10.3758/BRM.42.4.992).